Enhancing a Rule-Based MT System with Cross-Lingual WSD

Rudnick, Alex ; Rios, Annette ; Gasser, Michael

Abstract: Lexical ambiguity is a significant problem facing rule-based machine translation systems, as many words have several possible translations in a given target language, each of which can be considered a sense of the word from the source language. The difficulty of resolving these ambiguities is mitigated for statistical machine translation systems for language pairs with large bilingual corpora, as large n-gram language models and phrase tables containing common multi-word expressions can encourage coherent word choices. For most language pairs these resources are not available, so a primarily rule-based approach becomes attractive. In cases where some training data is available, though, we can investigate hybrid RBMT and machine learning approaches, leveraging small and potentially growing bilingual corpora. In this paper we describe the integration of statistical cross-lingual word-sense disambiguation software with SQUOIA, an existing rule-based MT system for the Spanish-Quechua language pair, and show how it allows us to learn from the available bitext to make better lexical choices, with very few code changes to the base system. We also describe Chipa, the new open source CL-WSD software used for these experiments.

Originally published at:
Rudnick, Alex; Rios, Annette; Gasser, Michael (2014). Enhancing a Rule-Based MT System with Cross-Lingual WSD. In: SaLTMiL Workshop on free/open-source language resources for the machine translation of less-resourced languages (LREC'14), Reykjavik, Iceland, 22 May 2014. SALTMIL, 31-36.

Enhancing a Rule-Based MT System with Cross-Lingual WSD

Alex Rudnick[1], Annette Rios[2], Michael Gasser[1]
[1] School of Informatics and Computing, Indiana University
[2] Institute of Computational Linguistics, University of Zurich
{alexr,gasser}@indiana.edu, arios@ifi.uzh.ch

Abstract

Lexical ambiguity is a significant problem facing rule-based machine translation systems, as many words have several possible translations in a given target language, each of which can be considered a sense of the word from the source language. The difficulty of resolving these ambiguities is mitigated for statistical machine translation systems for language pairs with large bilingual corpora, as large n-gram language models and phrase tables containing common multi-word expressions can encourage coherent word choices. For most language pairs these resources are not available, so a primarily rule-based approach becomes attractive. In cases where some training data is available, though, we can investigate hybrid RBMT and machine learning approaches, leveraging small and potentially growing bilingual corpora. In this paper we describe the integration of statistical cross-lingual word-sense disambiguation software with SQUOIA, an existing rule-based MT system for the Spanish-Quechua language pair, and show how it allows us to learn from the available bitext to make better lexical choices, with very few code changes to the base system. We also describe Chipa, the new open source CL-WSD software used for these experiments.

Keywords: under-resourced languages, hybrid machine translation, word-sense disambiguation

1. Introduction

Here we report on the development of Chipa, a package for statistical lexical selection, and on integrating it into SQUOIA,[1] a primarily rule-based machine translation system for the Spanish-Quechua language pair. With very few code changes to SQUOIA, we were able to make use of the lexical suggestions provided by Chipa.

The integration enables SQUOIA to take advantage of any available bitext without significantly changing its design, and to improve its word choices as additional bitext becomes available. Our initial experiments also suggest that we are able to use unsupervised approaches on monolingual Spanish text to further improve results.

In this paper, we describe the designs of the Chipa and SQUOIA systems, discuss the data sets used, and give results on both how well Chipa is able to learn lexical selection classifiers in isolation, and to what extent it is able to improve the output of SQUOIA on a full Spanish-to-Quechua translation task.

In its current design, SQUOIA makes word choices based on its bilingual lexicon; the possible translations for a given word or multi-word expression are retrieved from a dictionary on demand. If there are several possible translations for a lexical item, these are passed along the pipeline so that later stages can make a decision, but if the ambiguity persists, then the first entry retrieved from the lexicon is selected. While there are some rules for lexical selection, they have been written by hand and only cover a small subset of the vocabulary in a limited number of contexts.

In this work, we supplement these rules with classifiers learned from Spanish-Quechua bitext. These classifiers make use of regularities that may not be obvious to human rule-writers, providing improved lexical selection for any word type that has adequate coverage in the training corpus.

Quechua is a group of closely related indigenous American languages spoken in South America. There are many dialects of Quechua; SQUOIA focuses on the Cuzco dialect, spoken around the Peruvian city of Cuzco. Cuzco Quechua has about 1.5 million speakers and some useful available linguistic resources, including a small treebank (Rios et al., 2009), also produced by the SQUOIA team.

2. SQUOIA

SQUOIA is a deep-transfer RBMT system based on the architecture of MATXIN (Alegria et al., 2005; Mayor et al., 2011). The core system relies on a classical transfer approach and is mostly rule-based, with a few components based on machine learning. SQUOIA uses a pipeline approach, both in an abstract architectural sense and in the sense that its pieces are instantiated as a series of scripts that communicate via UNIX pipes. Each module performs some transformation on its input and passes along the updated version to the next stage. Many modules focus on very particular parts of the representation, leaving most of their input unchanged.

In the first stages, Spanish source sentences are analyzed with off-the-shelf open-source NLP tools. To analyze the input Spanish text, SQUOIA uses FreeLing (Padró and Stanilovsky, 2012) for morphological analysis and named-entity recognition, Wapiti (Lavergne et al., 2010) for tagging, and DeSr (Attardi et al., 2007) for parsing. All of these modules rely on statistical models.

In the next step, the Spanish verbs must be disambiguated in order to assign them a Quechua verb form for generation: a rule-based module tries to assign a verb form to each verb chunk based on contextual information. If the rules fail to do so due to parsing or tagging errors, the verb is marked as ambiguous and passed on to an SVM classifier, which assigns a verb form even if the context of that verb does not unambiguously select a target form. This is among the most difficult parts of the translation process,

[1] http://code.google.com/p/squoia/

as the grammatical categories encoded in verbs differ substantially between Spanish and Quechua. In the next step, a lexical transfer module inserts all possible translations for every word from a bilingual dictionary. Then a set of rules disambiguates the forms with lexical or morphological ambiguities. However, this rule-based lexical disambiguation is very limited, as it is not feasible to cover all possible contexts for every ambiguous word with rules.
The rest of the system makes use of a classical transfer procedure. A following module moves syntactic information between the nodes and the chunks in the tree, and finally, the tree is reordered according to the basic word order in the target language. In the last step, the Quechua surface forms are morphologically generated through a finite state transducer.

3. CL-WSD with Chipa

Chipa is a system for cross-lingual word sense disambiguation (CL-WSD). [2] By CL-WSD, we mean the problem of assigning labels to polysemous words in source-language text, where each label is a word or phrase type in the target language.
This framing of word-sense disambiguation, in which we consider the possible senses of a source-language word to be its known target-language translations, neatly addresses the problem of choosing an appropriate sense inventory, which has historically been a difficult problem for the practical application of WSD systems (Agirre and Edmonds, 2006). Here the sense distinctions that the CL-WSD system should learn are exactly those that are lexicalized in the target language. The CL-WSD framing also sidesteps the "knowledge acquisition bottleneck" hampering other work in WSD (Lefever et al., 2011). While supervised CL-WSD methods typically require bitext for training, this is more readily available than the sense-annotated text that would otherwise be required.
To appreciate the word-sense disambiguation problem embedded in machine translation, consider for a moment the different senses of "have" in English. In *have a sandwich*, *have a bath*, *have an argument*, and even *have a good argument*, the meaning of the verb "to have" is quite different. It would be surprising if our target language, especially if it is not closely related, used a light verb that could appear in all of these contexts.
A concrete example for different lexicalization patterns in Spanish and Quechua are the transitive motion verbs: The Spanish lemmas contain information about the path of the movement, e.g. *traer* - 'bring (here)' vs. *llevar* - 'take (there)'. Quechua roots, on the other hand, use a suffix (-*mu*) to express direction, but instead lexicalize information about the manner of movement and the object that is being moved. Consider the following examples:

general motion verbs:

- *pusa-(mu-)*: 'take/bring a person'
- *apa-(mu-)-*: 'take/bring an animal or an inanimated object'

motion verbs with manner:

- *marq'a-(mu-)*: 'take/bring smth. in one's arms'
- *q'ipi-(mu-)*: 'take/bring smth. on one's back or in a bundle'
- *millqa-(mu-)*: 'take/bring smth. in one's skirts'
- *hapt'a-(mu-)*: 'take/bring smth. in one's fists'
- *lluk'i-(mu-)*: 'take/bring smth. below their arms'
- *rikra-(mu-)*: 'take/bring smth. on one's shoulders'
- *rampa-(mu-)*: 'take/bring a person holding their hand'

The correct translation of Spanish *traer* or *llevar* into Quechua thus depends on the context. Furthermore, different languages simply make different distinctions about the world. The Spanish *hermano* 'brother', *hijo* 'son' and *hija* 'daughter' all translate to different Quechua terms based on the person related to the referent; a daughter relative to her father is *ususi*, but when described relative to her mother, *warmi wawa* (Academia Mayor de La Lengua Quechua, 2005).
Chipa, then, must learn to make these distinctions automatically, learning from examples in available word-aligned bitext corpora. Given such a corpus, we can discover the different possible translations for each source-language word, and with supervised learning, how to discriminate between them. Since instances of a source-language word may be NULL-aligned, both in the training data and in actual translations, we allow users to request classifiers that consider NULL as a valid label for classification, or not, as appropriate for the application.
The software holds all of the available bitext in a database, retrieving the relevant training sentences and learning classifiers on demand. If a source word has been seen with multiple different translations, then a classifier will be trained for it. If it has been seen aligned to only one target-language type, then this is simply noted, and if the source word is not present in the training data, then that word is marked out-of-vocabulary. Memory permitting, these classifiers and annotations are kept cached for later usage. Chipa can be run as a server, providing an interface whereby client programs can request CL-WSD decisions over RPC.
Here classifiers are trained with the scikit-learn machine learning package (Pedregosa et al., 2011), using logistic regression (also known as "maximum entropy") with the default settings and the regularization constant set to $C = 0.1$. We also use various utility functions from NLTK (Bird et al., 2009).
For this work, we use familiar features for text classification: the surrounding lemmas for the current token (three on either side) and the bag-of-words features for the entire current sentence. We additionally include, optionally, the Brown cluster labels (see below for an explanation), both for the immediate surrounding context and the entire sentence. We suspect that more feature engineering, particularly making use of syntactic information and surface word forms, will be helpful in the future.

[2] Chipa the software is named for chipa the snack food, popular in many parts of South America. It is a cheesy bread made from cassava flour, often served in a bagel-like shape in Paraguay. Also *chipa* means 'rivet, bolt, screw' in Quechua, something for holding things together. The software is available at `http://github.com/alexrudnick/chipa` under the GPL.

- lemmas from surrounding context (three tokens on either side)
- bag of lemmas from the entire sentence
- Brown cluster labels from surrounding context
- bag of Brown cluster labels from the entire sentence

Figure 1: Features used in classification

3.1. System Integration

In order to integrate Chipa into SQUOIA, we added an additional lexical selection stage to the SQUOIA pipeline, occurring after the rule-based disambiguation modules. This new module connects to the Chipa server to request translation suggestions – possibly several per word, ranked by their probability estimates – then looks for words that SQUOIA currently has marked as ambiguous.

For each word with multiple translation possibilities, we consider each of the translations known to SQUOIA and take the one ranked most highly in the results from the classifiers. If there are no such overlapping translations, we take the default entry suggested by SQUOIA's dictionary. Notably, since Chipa and SQUOIA do not share the same lexicon and bitext alignments may be noisy, translations observed in the bitext may be unknown to the SQUOIA system, and lexical entries in the SQUOIA dictionary may not be attested in the training data.

3.2. Learning From Monolingual Data

While in this work, our target language is under-resourced, we have many language resources available for the source language. We would like to use these to make better sense of the input text, giving our classifiers clearer signals for lexical selection in the target language.

One resource for Spanish is its abundant monolingual text. Given large amounts of Spanish-language text, we can use unsupervised methods to discover semantic regularities. In this work we apply Brown clustering (Brown et al., 1992), which has been used successfully in a variety of text classification tasks (Turian et al., 2010) and provides a straightforward mechanism to add features learned from monolingual text.

The Brown clustering algorithm takes as input unannotated text and produces a mapping from word types in that text to clusters, such that words in the same cluster have similar usage patterns according the corpus's bigram statistics. We can then use this mapping from words to clusters in our classifiers, adding an additional annotation for each word that allow the classifiers to find higher-level abstractions than surface-level words or particular lemmas. The desired number of clusters must be set ahead of time, but is a tunable parameter. We use a popular open source implementation of Brown clustering, [3] described by Liang (2005), running on both the Spanish side of our bitext corpus and on the Europarl corpus (Koehn, 2005) for Spanish.

[3] `https://github.com/percyliang/brown-cluster`

Figure 2 shows some illustrative examples of clusters that we found in the Spanish Europarl corpus. Examining the output of the clustering algorithm, we see some intuitively satisfying results; there are clusters corresponding to the names of many countries, some nouns referring to people, and common transitive verbs. Note that the clustering is unsupervised, and the labels given are not produced by the algorithm.

4. Experiments

Here we report on two basic experimental setups, including an *in-vitro* evaluation of the CL-WSD classifiers themselves and an *in-vivo* experiment in which we evaluate the translations produced by the SQUOIA system with the integrated CL-WSD system.

4.1. Classification Evaluation

To evaluate the classifiers in isolation, we produced a small Spanish-Quechua bitext corpus from a variety of sources, including the Bible, some government documents such as the constitution of Peru and several short folktales and works of fiction. The great majority of this text was the Bible. We used Robert Moore's sentence aligner (Moore, 2002), with the default settings to get sentence-aligned text. Initially there were just over 50 thousand sentences; 28,549 were included after sentence alignment.

During preprocessing, Spanish multi-word expressions identifiable with FreeLing were replaced with special tokens to mark that particular expression, and both the Spanish and Quechua text were lemmatized. We then performed word-level alignments on the remaining sentences with the Berkeley aligner (DeNero and Klein, 2007), resulting in one-to-many alignments such that each Spanish word is aligned to zero or more Quechua words, resulting in a label for every Spanish token.

With this word-aligned bitext, we can then train and evaluate classifiers. We evaluate here classifiers for the 100 most common Spanish lemmas appearing in the aligned corpus. For this test, we performed 10-fold cross-validation for each lemma, retrieving all of the instances of that lemma in the corpus, extracting the appropriate features, training classifiers, then testing on that held-out fold.

We report on two different scenarios for the *in-vitro* setting; in one case, we consider classification problems in which the word in question may be aligned to NULL, and in the other setting, we exclude NULL alignments. While the former case will be relevant for other translation systems, in the architecture of SQUOIA, lexical selection modules may not make the decision to drop a word. In both cases, we show the average classification accuracy across all words and folds, weighted by the size of each test set.

Here we compare the trained classifiers against the "most-frequent sense" (MFS) baseline, which in this setting is the most common translation for a given lemma, as observed in the training data.

We additionally show the effects on classification accuracy of adding features derived from Brown clusters, with clusters extracted from both the Europarl corpus and the Spanish side of our training data. We tried several different settings for the number of clusters, ranging from $C = 100$ to

category	top twenty word types by frequency
countries	francia irlanda alemania grecia italia españa rumanía portugal polonia suecia bulgaria austria finlandia hungría bélgica japón gran_bretaña dinamarca luxemburgo bosnia
more places	kosovo internet bruselas áfrica iraq lisboa chipre afganistán estrasburgo oriente_próximo copenhague asia chechenia gaza oriente_medio birmania londres irlanda_del_norte berlín barcelona
mostly people	hombre periodista jefes_de_estado individuo profesor soldado abogado delincuente demócrata dictador iglesia alumno adolescente perro chico economista gato jurista caballero bebé
infrastructure	infraestructura vehículo buque servicio_público cultivo edificio barco negocio motor avión monopolio planta ruta coche libro aparato tren billete actividad_económica camión
common verbs	pagar comprar vender explotar practicar soportar exportar comer consumir suministrar sacrificar fabricar gobernar comercializar cultivar fumar capturar almacenar curar beber

Figure 2: Some illustrative clusters found by the Brown clustering algorithm on the Spanish Europarl data. These are five out of $C = 1000$ clusters, and were picked and labeled arbitrarily by the authors. The words listed are the top twenty terms from that cluster, by frequency.

system	accuracy				
MFS baseline	54.54				
chipa, only word features	65.43				
	$C = 100$	$C = 200$	$C = 500$	$C = 1000$	$C = 2000$
chipa, +clusters from training bitext	66.71	67.43	68.41	69.00	69.43
chipa, +clusters from europarl	66.60	67.18	67.83	68.25	68.58

Figure 3: Results for the *in-vitro* experiment; classification accuracies over tenfold cross-validation including null-aligned tokens, as percentages.

system	accuracy				
MFS baseline	53.94				
chipa, only word features	68.99				
	$C = 100$	$C = 200$	$C = 500$	$C = 1000$	$C = 2000$
chipa, +clusters from training bitext	71.53	72.62	73.88	74.29	74.78
chipa, +clusters from europarl	71.27	72.08	73.04	73.52	73.83

Figure 4: Classification accuracies over tenfold cross-validation, excluding null-aligned tokens.

$C = 2000$. In all of our experimental settings, the addition of Brown cluster features substantially improved classification accuracy. We note a consistent upward trend in performance as we increase the number of clusters, allowing the clustering algorithm to learn finer-grained distinctions. The training algorithm takes time quadratic in the number of clusters, which becomes prohibitive fairly quickly, so even finer-grained distinctions may be helpful, but will be left to future work. On a modern Linux workstation, clustering Europarl (2M sentences) into 2000 clusters took roughly a day.

The classifiers using clusters extracted from the Spanish side of our bitext consistently outperformed those learned from the Europarl corpus. We had an intuition that the much larger corpus (nearly two million sentences) would help, but the clusters learned in-domain, largely from the Bible, reflect usage distinctions in that domain. Here we are in fact cheating slightly, as information from the complete corpus is used to classify parts of that corpus.

Figures 3 and 4 show summarized results of these first two experiments.

4.2. Translation Evaluation

In order to evaluate the effect of Chipa on lexical selection in a live translation task, we used SQUOIA to translate two Spanish passages for which we had reference Quechua translations. The first is simply a thousand sentences from the Bible; the second is adapted from the Peruvian government's public advocacy website,[4] which is bilingual and presumably contains native-quality Quechua. We collected and hand-aligned thirty-five sentences from this site.

Having prepared sentence-aligned and segmented bitexts for the evaluation, we then translated the Spanish side with SQUOIA, with various CL-WSD settings to produce Quechua text. In comparing the output Quechua with the reference translations, BLEU scores were quite low. The output often contained no 4-grams that matched with the reference translations, resulting in a geometric mean of 0. So here we report on the unigram-BLEU scores, which reflect some small improvements in lexical choice. See Figure 5 for the numerical results.

On the web test set, unfortunately very few of the Spanish

[4] *Defensoría del Pueblo*, `http://www.defensoria.gob.pe/quechua.php`

system	web test set	bible test set
squoia without CL-WSD	28.1	24.2
squoia+chipa, only word features	28.1	24.5
squoia+chipa, +europarl clusters	28.1	24.5
squoia+chipa, +bible clusters	28.1	24.5

Figure 5: BLEU-1 scores (modified unigram precision) for the various CL-WSD settings of SQUOIA on the two different Spanish-Quechua test sets.

words used were both considered ambiguous by SQUOIA's lexicon and attested in our training corpus. Enabling Chipa during translation, classifiers are only called on six of the thirty-five sentences, and then the classifiers only disagree with the default entry from the lexicon in one case.

We do see a slight improvement in lexical selection when enabling Chipa on the Bible test set; the three feature settings listed actually all produce different translation output, but they are of equal quality. Here the in-domain training data allowed the classifiers to be used more often; 736 of the thousand sentences were influenced by the classifiers in this test set.

5. Related Work

Framing the resolution of lexical ambiguities in machine translation as an explicit classification task has a long history, dating back at least to early SMT work at IBM (Brown et al., 1991). More recently, Carpuat and Wu have shown how to use classifiers to improve modern phrase-based SMT systems (Carpuat and Wu, 2007). CL-WSD has received enough attention to warrant shared tasks at recent SemEval workshops; the most recent running of the task is described by Lefever and Hoste (2013). In this task, participants are asked to translate twenty different polysemous English nouns into five different European languages, in a variety of contexts.

Lefever *et al.*, in work on the ParaSense system (2011), produced top results for this task with classifiers trained on local contextual features, with the addition of a bag-of-words model of the translation of the complete source sentence into other (neither the source nor the target) languages. At training time, the foreign bag-of-words features for a sentence are extracted from available parallel corpora, but at testing time, they must be estimated with a third-party MT system, as they are not known a priori. This work has not yet, to our knowledge, been integrated into an MT system on its own.

In our earlier work, we prototyped a system that addresses some of the issues with ParaSense, requiring more modest software infrastructure for feature extraction while still allowing CL-WSD systems to make use of several mutually parallel bitexts that share a source language (Rudnick et al., 2013). We have also done some previous work on CL-WSD for translating into indigenous American languages; an earlier version of Chipa, for Spanish-Guarani, made use of sequence models to jointly predict all of the translations for a sentence at once (Rudnick and Gasser, 2013).

Francis Tyers, in his dissertation work (2013), provides an overview of lexical selection systems and describes methods for learning lexical selection rules based on available parallel corpora. These rules make reference to the lexical items and parts of speech surrounding the word to be translated. Once learned, these rules are intended to be understandable and modifiable by human language experts. For practical use in the Apertium machine translation system, they are compiled to finite-state transducers.

Rios and Göhring (2013) describe earlier work on extending the SQUOIA MT system with machine learning modules. They used classifiers to predict the target forms of verbs in cases where the system's hand-crafted rules cannot make a decision based on the current context.

6. Conclusions and Future Work

We have described the Chipa CL-WSD system and its integration into SQUOIA, a machine translation system for Spanish-Quechua. Until this work, SQUOIA's lexical choices were based on a small number of hand-written lexical selection rules, or the default entries in a bilingual dictionary.

We have provided a means by which the system can make some use of the available training data, both bilingual and monolingual, with very few changes to SQUOIA itself. We have also shown how Brown clusters, either when learned from a large out-of-domain corpus or from a smaller in-domain corpus, provide useful features for a CL-WSD task, substantially improving classification accuracy.

In order make better use of the suggestions from the CL-WSD module, we may need to expand the lexicon used by the translation system, so that mismatches between the vocabulary of the available bitext, the translation system itself, and the input source text do not hamper our efforts at improved lexical selection. Finding more and larger sources of bitext for this language pair would of course help immensely.

We would like to learn from the large amount of monolingual Spanish text available; while the Europarl corpus is nontrivial, there are much larger sources of Spanish text, such as the Spanish-language Wikipedia. We plan to apply more clustering approaches and other word-sense discrimination techniques to these resources, which will hopefully further improve CL-WSD across broader domains.

Better feature engineering outside of unsupervised clusters may also be useful. In the future we we will extract features from the already-available POS tags and the syntactic structure of the input sentence.

We also plan to apply the Chipa system to other machine translation systems and other language pairs, especially Spanish-Guarani, another important language pair for South America.

7. References

Academia Mayor de La Lengua Quechua. (2005). *Diccionario: Quechua - Español - Quechua, Qheswa - Español - Qheswa: Simi Taqe, 2da ed.* Cusco, Perú.

Agirre, E. and Edmonds, P. G. (2006). *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science+ Business Media.

Alegria, I., de Ilarraza, A. D., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., Forcada, M. L., Rojas, S. O., and Padró, L. (2005). An Open Architecture for Transfer-based Machine Translation between Spanish and Basque. In *Workshop on Open-source machine translation at Machine Translation Summit X*.

Attardi, G., Dell'Orletta, F., Simi, M., Chanev, A., and Ciaramita, M. (2007). Multilingual dependency parsing and domain adaptation using DeSR. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1112–1118.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1991). Word-Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Carpuat, M. and Wu, D. (2007). How Phrase Sense Disambiguation Outperforms Word Sense Disambiguation for Statistical Machine Translation. In *11th Conference on Theoretical and Methodological Issues in Machine Translation*.

DeNero, J. and Klein, D. (2007). Tailoring Word Alignments to Syntactic Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, June.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of The Tenth Machine Translation Summit*.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lefever, E. and Hoste, V. (2013). SemEval-2013 Task 10: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Lefever, E., Hoste, V., and De Cock, M. (2011). ParaSense or How to Use Parallel Corpora for Word Sense Disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Liang, P. (2005). Semi-supervised learning for natural language. Master's thesis, MIT.

Mayor, A., Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., and Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for basque. *Machine Translation*, 25(1):53–82.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *AMTA*, pages 135–144.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rios, A., Göhring, A., and Volk, M. (2009). A quechua–spanish parallel treebank. In *Proceedings of 7th Workshop on Treebanks and Linguistic Theories (TLT-7), Groningen*.

Rios Gonzales, A. and Göhring, A. (2013). Machine Learning Disambiguation of Quechua Verb Morphology. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, Sofia, Bulgaria.

Rudnick, A. and Gasser, M. (2013). Lexical Selection for Hybrid MT with Sequence Labeling. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 102–108, Sofia, Bulgaria.

Rudnick, A., Liu, C., and Gasser, M. (2013). HLTDI: CL-WSD Using Markov Random Fields for SemEval-2013 Task 10. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

Turian, J., Ratinov, L.-A., and Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden.

Tyers, F. M. (2013). *Feasible lexical selection for rule-based machine translation*. Ph.D. thesis, Departament de Llenguatges i Sistemes Infomàtics, Universitat d'Alacant.